SLAC-PUB-10026
July 2003

Interfacing Interactive Data Analysis Tools with the Grid: The PPDG CS-11 Activity*

D. L. Olson[a]J. Perl[b]

[a]Lawrence Berkeley National Laboratory,
Berkeley, CA, 94720, USA

[b]Stanford Linear Accelerator Center,
Menlo Park, CA, 94025, USA

For today's physicists, who work in large geographically distributed collaborations, the data grid promises significantly greater capabilities for analysis of experimental data and production of physics results than is possible with today's "remote access" technologies. The goal of letting scientists at their home institutions interact with and analyze data as if they were physically present at the major laboratory that houses their detector and computer center has yet to be accomplished. The Particle Physics Data Grid project (www.ppdg.net) has recently embarked on an effort to "Interface and Integrate Interactive Data Analysis Tools with the grid and identify Common Components and Services." The initial activities are to collect known and identify new requirements for grid services and analysis tools from a range of current and future experiments (ALICE, ATLAS, BaBar, D0, CMS, JLab, STAR, others welcome), to determine if existing plans for tools and services meet these requirements. Follow-on activities will foster the interaction between grid service developers, analysis tool developers, experiment analysis framework developers and end user physicists, and will identify and carry out specific development/integration work so that interactive analysis tools utilizing grid services actually provide the capabilities that users need. This talk will summarize what we know of requirements for analysis tools and grid services, as well as describe the identified areas where more development work is needed.

1. Introduction

While considerable work has been done to bring batch production processing of physics experiments onto the grid, very little has been done specifically with the point of view of interactive data analysis. Within the Particle Physics Data Grid project (www.ppdg.net) an activity to identify and address the relevant issues was recently started under the label CS-11.

Given the large scope of possibilities with the grid we think it is useful to describe four models of how interactive data analysis could be carried out using grid services. Even though these model vary widely in apparent complexity, there is significant commonality in what grid services need to provide. For brevity here we just label these models

- grid as black box
- real-time batch
- interactive batch
- pre-started analysis services

and refer the reader to a longer document that gives more detail.[2]

We compiled a number of use cases that comprise the typical steps that a physicists takes in carrying out interactive data analysis, and these are described in the next section. These use cases were considered by a group of people at a workshop in Berkeley, California in June 2002. [1].

*This work is supported by the U.S. Department of Energy Office of Science, the Office of High Energy and Nuclear Physics and the Office of Advanced Scientific Computing Research under contract Nos. DE-AC03-76SF00098 and DE-AC03-76SF00515

2. Steps in data analysis (use cases) considered

In the subsections below a brief description is given describing the use case activity and the reader is refered to [2] for additional detail.

2.1. Example: Select Data

Users start by specifying what data is to be analyzed.

Data Catalog Service:

1. must be very reliable, easy to access, queryable (accessible to end user and to applications)
2. should have a hierarchical view (may have different overlapping views), like a file system (able to browse, set access controls, highly reliable, allow locks to protect from deletion).
3. must include quality of service information (how accessible in location and/or time is the data, what is its proximity to processing power)
4. must include information at the event component level,not just the event level.
5. must include attributes to identify various possible characteristics of "good" events.
6. must include a way to tell what calibration/geometry databases are associated with a given set of data.
7. must allow the user to create persistent handles to denote the data they have selected (and associated calibration/geometry databases).

Job control system or interactive analysis interface:

1. must be able to accept data selection via persistent handle created during data selection process.
2. must be able to scale to a global set of resources with priority, policy and quota limits. The complexity of job control, that arises from the constraints, are much more evident in the case of interactive analysis.

2.2. Example: Get Smaller Local Subset of Data

Replica management services:

1. should allow user to move selected smaller subset of data to their local machine.
2. may need to merge small amounts of data from several different locations to create the desired smaller subset of data.
3. may need to be able to formulate data delivery strategy: timing and/or scheduling data delivery, optimization of choice of data sources.

2.3. Remaining use cases

The remaining use cases considered are listed below and the reader can find more description in [2].

- Inspect Smaller Dataset to Develop Cuts and Analysis Programs
- Move Data
- Choose Standard or Modified Versions of Code
- Run analysis on small data subset
- Retrieve results
- Estimate resources for larger job
- Negotiate availability or access to resources
- Run analysis on large data set
- Check status of analysis in progress
- View results of analysis in progress
- Suspend/resume analysis in progress
- Abort analysis in progress
- View results of completed analysis
- Displays of selected individual events at varying levels of detail
- Add refined data to data store
- Share refined data with collaborators
- Add tag data
- Compare results
- Calculate cross sections
- Maintain audit trail (data provenance)

3. Conclusions and Outstanding Issues

While the working physicist will find few surprises in the above requirements, it is clear that interactive analysis places significant requirements on grid services.

Grid services most strongly impacted by interactive analysis requirements are the Data Catalog Service and the Job Control System (or Interactive Analysis Interface), but many other services are also involved. The heavy requirements placed on the Data Catalog Service strongly suggest the use of a Database Management System behind this service.

3.1. Outstanding Issues

Some of the outstanding issues are listed below.

3.1.1. Need for standard data definitions

At least among the group assembled for the CS11 workshop at Berkeley in June 2002, there was significant confusion over data definition terminology (such as "dataset"). Agreement on clearer definitions must be made for documents such as this to be effective. It may be best to take these definitions from the HEPCAL document[3].

3.1.2. Need for "grid object" definition

Interactive analysis users are generally concerned with data at a finer grain than just whole events. They need to know which components of a given event are available for their analysis (TAG, AOD, ESD, SIM, etc.). To make progress in this area, it would be useful to develop a shared "grid object" definition.

3.1.3. Need for debugging tools appropriate for the grid environment

New code is often used during interactive analysis. If this analysis is to proceed in a grid environment, distributed debugging tools must be available to help the user. It is not sufficient for the user to debug their code just on their local machine. They must have tools to help with more difficult questions such as why their code that ran fine on their local machine then fails on some other grid node.

3.1.4. Need for common metadata catalog schema

If different experiments could agree on a common metadata catalog schema, it would make it much easier to design analysis tools that can be shared across experiments.

3.1.5. Matchmaking services

Will matchmaking services be so smart that they can tell the analysis tool where to run, or will they simply return a list of nodes (cross-referenced by cpu and data access ability) that serves as input to decision making by the desktop tool or portal?

3.1.6. User interface portal

Understand requirements for user view of grid services and interactions relevant to interactive data analysis. Explore applicable technology and prototype grid-enabled interactive analysis environments.

REFERENCES

1. http://www.ppdg.net/mtgs/18jun02-lbl/agenda.htm.
2. "Grid Service Requirements for Interactive Data Analysis", http://www.ppdg.net/pa/ppdg-pa/idat/analysis_use_cases-grid-reqs.pdf
3. http://fca.home.cern.ch/fca/HEPCAL.doc